MARKER DETECTION OF CONTRACT CHEATING: AN INVESTIGATIVE CORPUS LINGUISTIC APPROACH

Olumide Popoola[1]

[1] *Queen Mary University of London, United Kingdom*

INTRODUCTION

An increased focus on detection rather than prevention of contract cheating [1] has placed assessment markers in the frontline to preserve academic integrity. Consequently, tools are needed that can increase detection during the marking process. Text-based approaches have shown potential. [2] demonstrated that marker detection efforts can be improved through exposure to linguistic reports generated by Turnitin's Authorship Investigate software; [3] demonstrates that stylometric analysis can be used to verify authorship. Whilst both tools can provide further evidence after suspicions have already been raised by an individual student submission, such tools are not designed for use during routine marking.

What if commercial essay writing has distinctive linguistic features? Markers could look for signs of commercial essay writing while marking; assessments could potentially be designed to hinder commercial essay writers. In this paper, a multi-discipline analysis of student and commercial essays, using the most comprehensive set of linguistic features deployed in academic integrity research to date, provides proof-of-concept for linguistics-based detection of outsourced writing

Linguistics-based approaches have been used to detect deception and disinformation in online news, consumer reviews and social media. Commercial essay writing is a form of deception comparable to fake review writing; both use 'gig economy' professional writers recruited through third-party websites. This research deploys investigative corpus linguistic techniques used in the detection of fake news and fake online reviews.

RESEARCH QUESTION AND HYPOTHESIS

Specifically this paper presentation tests the following hypotheses:

- that commercial and student essays will differ systematically on a range of linguistic features.
- that a predictive model can be built to classify student and commercial texts at a rate significantly above chance.

DATA

Linguistic deception detection uses text classification to build predictive statistical models trained on text data labelled for veracity. Clever data collection is key to the investigative corpus linguistic approach. Purchasing sufficient essays to build a text classification model, whilst replicating the way students engage with these third-party websites, would be limited by financial and ethical issues. Instead, this research uses investigative corpus linguistic

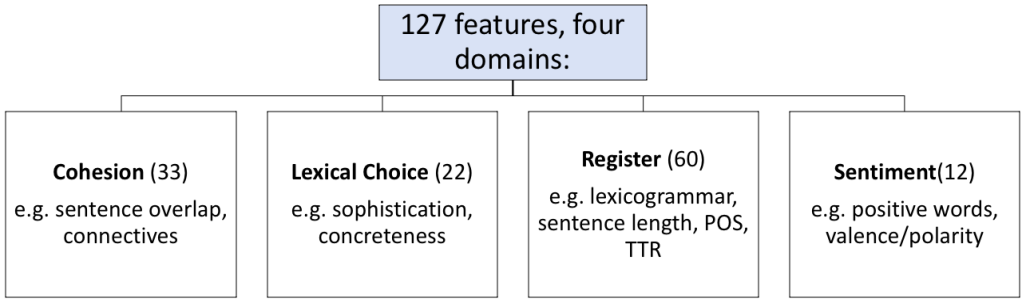

Fig. 1: Linguistic categories and example features

techniques to compile the 'Cheat-AI' corpus of commercial and student essays.

investigative techniques are characterised by their use of real-world data. The essays in this research were harvested from the internet using the Bootstrapping Corpora and Terms technique [4]. This process involves iteratively querying search engines with seed terms designed to find the required data. Investigative research identified phrases such as "expert writer" "sample essay" "plagiarism free essay" "student essay" as well as names of popular third- party websites as productive terms for finding student and commercial essays. Although commercial essays were far harder to find, a sufficient number were retrieved for discipline-level analysis. In total, 12347 student essays and 508 commercial essays were harvested in 30 subjects (Table 1).

Tab. 1: Cheat-AI corpus: 508 commercial essays

Subject	Number of essays
Business	79
Law	50
Nursing	45
Health	30
Education	25
Other Business Cognate disciplines	98
Other Humanities and Social Sciences	128
STEMM	54

This paper reports on the results of the application of this approach to the three most common commercial essay subjects: Business, Law and Nursing essays.

LINGUISTIC FEATURES

Significantly expanding the stylometric approach used for authorship analysis in [3]. 127 linguistic features were extracted in four domains to provide a comprehensive and holistic representation of the cognitive, functional and emotional aspects of the writing process (see Figure 1). The Suite of Automatic Linguistic Analysis Tools [5] was used to extract 67 features related to cohesion, lexical choice and sentiment; 60 features related to linguistic register and style were extracted using MAT Tagger [6].

These 127 features were then fed into a binary logistic regression with essay veracity as the dependent variable (Commercial = 1; Student = 0) in order to produce a predictive model. The model achieved 82% overall accuracy with a binary logistic regression text classification (Table 2).

To aid interpretation and facilitate assessment of the relative contribution of each domain to the model, Principal Components Analysis was conducted reduce the to identify components in the four domains separately. 30 components were detected across the four domains (Table 3). These components were also fed into a binary logistic regression; a loss of accuracy of less than 5% indicates that this set of components is a reliable representation of the linguistic data.

Tab. 2: Logistic Regression Classification Model (127 features)

	Predicted STUDENT	Predicted COMMERCIAL	%Correct
Actual STUDENT (1643)	1412	231	85.9%
Actual COMMERCIAL (698)	192	506	72.5%
Nagelke R^2=.513; Hosmer Lemshow = .362		**Overall percentage**	**81.9%**
		Majority class baseline	**70.0%**

Tab. 3: 30 linguistic factors in 4 categories identified by Principal Component Analysis

Component Categories → / Component number ↓	*Cohesion (Coh) (71.1% variance; KMO 0.68)*	*Lexical Choice (LexC) (79.2% variance; KMO= .60)*	*Sentiment (Sent) (82.5%; KMO=.51)*	*Register (Reg) (32.5% variance; KMO=.48)*
1	Lexicosemantic overlap (16.7%,)	Lexical Sophistication (26.7%)	Positive Emotion (45.6%)	Citation (6.6%)
2	Additive connectives (13.2%,	Lexical Diversity (14.3%)	Confident (16.1%)	Negative statements
3	Unspecified reference (12.6%,	Lexical Sparsity (11.8%)	Positive evaluation (8.6%)	Speculation 3.8%)
4	Reason/Logical Connectives (7.1%,	Lexical concreteness (9.9%)	Agitated (6.6%)	Description (predicative adjectives)(
5	Contradiction/Contrast (5.5%,	Semantic similarity (7.0%)	Positive events (5.8%)	Present tense (3.2%)
6	Disjunction/Negation (4.8%,	(4.9%) Sentence length		Informality
7	Causation (4.2%,	Lexical stance (4.6%)		(2.6%) Transitions
8	Temporal (3.6%,			(2.5%) Perfect aspect
9	Shell nouns 3.5%,			Subordination (2.4%)

Tab. 4: Linguistic features of commercial vs. student essay writing

Component description	Sig	Exp(B)	Component No.
Lexical Sophistication	.02	2.217	Lex1
Ambiguous reference	.03	1.695	Coh3
Transition	.03	1.605	Reg7
Shell nouns	.04	1.353	Coh9
Sparsity	.04	1.243	Lex3
Lexicosemantic overlap	.05	1.199	Coh1
Informality	.02	.489	Reg6
Negative statements	.03	.639	Coh4
Additive Connectives	.04	.753	Coh2
Positive evaluation	.04	.759	Sent3
Lexical Concreteness	.04	.764	Lex4

Commercial essay features

Student essay features

DISCUSSION

Commercial writing has a superficial quality – a conventional academic writing style and sophisticated vocabulary. It is also defective, due to its repetitiveness, high levels of redundancy and verbosity – all signs of a general padding strategy likely in response to the parameters of commercial writing such as word count and time constraints.

Specifically, the commercial writing features that generalised across Business, Law and Nursing essays were:

- Formal academic writing style (e.g. use of transitions, shell nouns).
- Combination of lexical sophistication and sparsity, indicating sesquipedalian prose style where writers sprinkle big words amongst circuitous language.
- Ambiguity due to unspecified reference words ('this', 'it')
- Repetition of content words and use of synonyms across adjacent sentences indicating sentence similarity and thesaurus use.

Markers could use the significant components as a checklist (Table 4) to flag suspicious submissions. The linguistic regression model could also be used in assessment security measures as an alternative to random sampling of cohort submissions.

REFERENCES

QAA (2020) *Contracting to Cheat in Higher Education* 2nd ed

Dawson, P, Sutherland-Smith, W and Ricksen, M. (2019) Can software improve marker accuracy at detecting contract cheating? A pilot study of the Turnitin authorship investigate alpha, *Assessment and evaluation in higher education*, pp. 1-10

Crockett, R. and Best, K. (2020) Stylometric Comparison of Professionally Ghost-Written and Student-Written Assignments. In *6th International Conference Plagiarism Across Europe and Beyond 2020*. Mendel University Press.

Baroni, M. and Bernardini, S. (2004) BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the International Conference on Language Resources and Evaluation* pp. 1313-1316

Crossley, S.A. and Kyle, K. (2018). Analyzing spoken and written discourse: A role for natural language processing tools. In *The Palgrave handbook of applied linguistics research methodology* (pp. 567-594). Palgrave Macmillan, London.

Nini, A. (2019). The Multi-Dimensional Analysis Tagger. In Berber Sardinha, T. and Veirano Pinto M. (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67-94, London; New York: Bloomsbury Academic